

DON'T BE EVIL:

A Perspective on the Rise and Ultimate Demise of Google



Peter Ubriaco
Rensselaer Polytechnic Institute, Troy, NY
Prepared for Intro to Communications Theory

November 12, 2004

Don't be evil is Google's unofficial motto¹. How this motto is applied within the company remains unclear. What is plain to see, however, is that Google is currently the hands down leader of Internet search traffic. What does it mean to rely so heavily on Google? How does Google affect the way that we obtain information? Is Google biased, and if so, how? Has Google become too big for its britches? If it is found that Google is in fact detrimental to free Internet communication, what can be done? Throughout this paper, these points will be discussed at length.

BACKGROUND

To fully understand the nature of Google, it is useful to have a grasp on the history of the Internet itself. The Internet is essentially an offshoot of one of the first computer networks ever built. Built by the United States Department of Defense, it was known as ARPAnet. By 1969, the Department of Defense had almost all but turned over control of ARPAnet to universities. Computer science graduate students eagerly poured their efforts at their particular schools to become a part of ARPAnet. During this same time period, the rise of the counterculture was also taking place. The same students that turned ARPAnet into a tool of any practical use in communication developed it in an open,

¹ <http://www.google.com/governance/conduct.html>, "Google Investor Relations," November 2004.

distributed fashion. As they communicated with each other electronically, they agreed to make the ARPAnet an "open club, that all were invited to enjoy."²

Over the next few decades, the Internet remained largely unknown. Essentially, it was useful only for email, Usenet newsgroups, or other raw data transmission. The Hypertext Transfer Protocol (HTTP), known affectionately as the World Wide Web protocol, was developed around 1991, for the purpose of distributing formatted data from a central server, available for retrieval at any time³ (it was initially developed to allow a group Physics professors to collaborate).

In 1993, a technical consulting firm by the name of Network Solutions, Inc. won a five year contract with the United States government to administer an organization called InterNIC, whose function was to administrate the issuance of ".com" and other such addresses⁴.

This could essentially be considered the birth of the Internet as it is known to many people today. For the casual web browser, there is little to Internet but what can be easily reached within Microsoft Internet Explorer. If this were not the case, services like America Online would be out of business, since essentially America Online is

² <http://www.historyoftheinternet.com/chap2.html>, "History of the Internet: ARPAnet is born," November 2004

³ http://www.hitmill.com/internet/web_history.asp, "History of the Web Beginning at CERN," November 2004

⁴ <http://www.cc.utah.edu/history.html>, "History of the Internet," November 2004

DON'T BE EVIL: A PERSPECTIVE ON THE RISE AND ULTIMATE DEMISE OF GOOGLE

an underperforming Internet Service Provider that loads their content with advertisements and affords little more than a single starting point from which to begin web browsing, email, and instant messaging. It is interesting to note that the same services could just as easily be provided without such incredible overhead or cost, but this is an entire topic altogether. The lesson to be learned from America Online is that of the power of the portal – the users that popularized America Online, which could essentially be considered the first true Internet superpower (and also the first casualty of the post-broadband *perestroika*) made it clear that Internet use should be simple and straightforward.

Throughout 1991 until 1994, usage of HTTP was increasing approximately ten-fold each year. This incredible rate of growth was generally unexpected decades before this. With this number of hosts connected together, there was an almost innumerable amount of information to be found, compared to the human capacity to actually take in that information. The question was: how can one find relevant information that they are seeking? Prior to the creation of HTTP, there were two other formats used for exchanging information: the File Transfer Protocol (FTP) and another known as Gopher, which could be described as a rudimentary text-only Internet. Yahoo's first generation web browser was

essentially a Gopher client⁵. The concept of these early engines was roughly the same as that of modern browsers: a set of HTTP (of FTP, or Gopher) documents are loaded into a spider program that systematically loads linked pages and related information and indexes it for later retrieval. Such a system made it possible for a user to obtain information without knowing a specific source for that particular knowledge. The concept was powerful, since it allowed relatively unknown information to be disseminated (theoretically) infinitely. Such a feat would be essentially impossible with other media such as newspapers, periodicals, and books. Further-more, information was accessible to anyone at any time. Most libraries could hardly claim to parallel such convenience and utility.

The Internet was not initially perceived to be a threat to traditional media because of its relative obscurity. When technologies like JPG, MPG, and MP3, used to conveniently transmit photos, movies, and music respectively, became popular, suddenly the other information outlets began to worry – with good cause.

To be able to simply enter a search term – be it a word, a phrase, a name, or any other moniker – and retrieve literally any published document anywhere in the globe truly illustrated the power

⁵<http://academ.hvcc.edu/~kantopet/misc/index.php?page=search+engines&parent=services>, "History: Search Engines," November 2004

DON'T BE EVIL: A PERSPECTIVE ON THE RISE AND ULTIMATE DEMISE OF GOOGLE

of interconnectivity. No radio, television, or print medium could possibly compete on the basis of speed. But as was mentioned earlier, the Internet was growing at an exponential rate at the same time that computers were effectively obeying Moore's law, causing the Internet to balloon at a rate much faster than that of the processing power available to effectively search it. It was clear that novel methods would need to be discovered to index such a breadth of material.

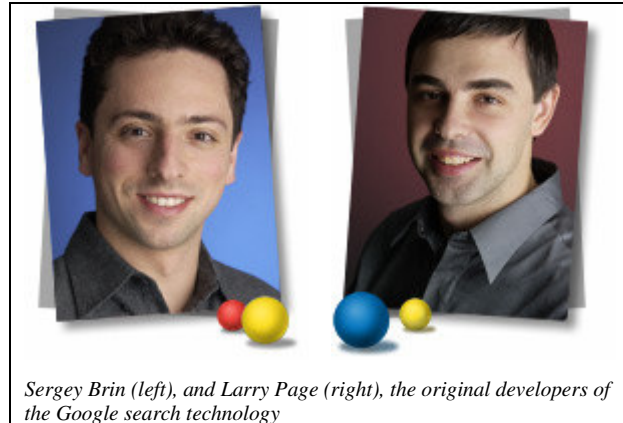
THE BIRTH OF GOOGLE

In 1995, Sergey Brin and Larry Page were two computer science students at Stanford that happened to make acquaintances. As it turned out, they were both researching methods to search and sort large datasets. At this time, the Internet was well established, but the boom that was to be the information age was still in its infancy.

Together they solved a burning question within the search technology community: how can this much data be parsed and sorted in such a way that results are returned based on relevance rather than some other arbitrary measure such as a tally of word occurrences? They developed software around their theories, and made their work pseudo-public through Stanford University under the name Google.

In 1998, they presented their technology at the World Wide Web Conference, and very soon

after were able to raise \$30 million in funding through venture capital, private investors, and financed the official public launch of Google in 1999.



At the time, their competition was quite diverse. A number of other companies such as Yahoo, Altavista, Lycos, Inktomi, Mirosoft, and America Online had invested significant amounts of effort into luring web users to their search portals, some more successfully than others. In 1999, Yahoo was considered the definitive search engine. It is useful to note that each of these search engines employed at least some degree of filtering technologies that attempted to find relevant documents, although their algorithms were incredibly simple and subsequently easily tampered with. A common phenomenon could be observed searching for terms such as 'sex' – because of the commercial interest in selling sex via the Internet, a number of third parties overwhelmed the simple web spiders through various techniques such as intentionally loading sites with text readable only by web spiders, creating multiple sites with the same content, and even exploiting other sites that incorporated

DON'T BE EVIL: A PERSPECTIVE ON THE RISE AND ULTIMATE DEMISE OF GOOGLE

automatic back linking to referring pages. This technique is also referred to as spamdexing⁶.

This technique is what would effectively define Google as the leader in Internet searching. While Yahoo and America Online were running archaic web spiders hardly capable of any intelligent indexing, Google appeared on the scene providing information with startling speed and accuracy. Additionally, Google was completely free to end users. At the time Google premiered, it did not incorporate any advertisements whatsoever – Google functioned *pro bono*. How could a service that operated at no cost to consumers be evil? It was the corporations such as Yahoo and Inktomi that had evil intentions of profiting via advertising revenue, or so many consumers seemed to believe.

Around 2001, Google had risen through the ranks to achieve status as the top search engine on the Internet⁷. Their unusually minimalist interface seemed inexplicably appealing to users, and carried the advantage of reducing server loads, bandwidth requirements, and end user load times. Recall that in 2001, dial-up Internet access was far more prominent than it is today, making the seemingly slight difference in page sizes significantly faster for users compared to Google's competitors. This, coupled with a

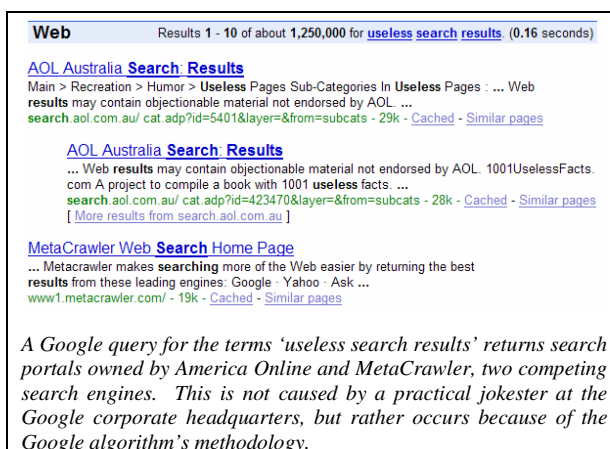
⁶ <http://en.wikipedia.org/wiki/Spamdexing>, "Spamdexing – Wikipedia, the free encyclopedia," November 2004

⁷ <http://en.wikipedia.org/wiki/Google>, "Google – Wikipedia, the free encyclopedia," November 2004

novel method of page indexing, made Google stand out from the crowd to which it belonged.

WHY IS GOOGLE SUCCESSFUL?

One simple technique that Google employed to rank pages was that of link popularity. Simply put, Google's algorithm, called PageRank, heavily weighted the proposition that more relevant sites will be linked to more often⁸. This ranking methodology assuredly must skew the results that are provided. The original intent of PageRank was to thwart efforts of so-called spamdexers, but of course was ultimately abused in different ways to the same effect. This technique is now known as Google bombing, and is generally used to generate unexpected, completely irrelevant (but often humorous) search results. Many people have previously assumed that Google tampered with its results by hand, but it seems that Google is in fact automatic as they suggest. As such, when one searches for "useless search results" and finds the



The screenshot shows a search engine interface with the following content:

- Header: **Web** Results 1 - 10 of about 1,250,000 for **useless search results**. (0.16 seconds)
- Result 1: [AOL Australia Search Results](#)
Main > Recreation > Humor > Useless Pages Sub-Categories In Useless Pages : ... Web results may contain objectionable material not endorsed by AOL. ...
search.aol.com.au/cat.adp?id=5401&layer=&from=subcats - 29k - [Cached](#) - [Similar pages](#)
- Result 2: [AOL Australia Search Results](#)
... Web results may contain objectionable material not endorsed by AOL. 1001UselessFacts.com A project to compile a book with 1001 useless facts. ...
search.aol.com.au/cat.adp?id=423470&layer=&from=subcats - 28k - [Cached](#) - [Similar pages](#)
[More results from search.aol.com.au]
- Result 3: [MetaCrawler Web Search Home Page](#)
... Metacrawler makes searching more of the Web easier by returning the best results from these leading engines: Google - Yahoo - Ask ...
www1.metacrawler.com/ - 19k - [Cached](#) - [Similar pages](#)

Below the search results, there is a paragraph: *A Google query for the terms 'useless search results' returns search portals owned by America Online and MetaCrawler, two competing search engines. This is not caused by a practical jokester at the Google corporate headquarters, but rather occurs because of the Google algorithm's methodology.*

⁸ <http://www.google.com/technology/>, "Google Technology," November 2004

DON'T BE EVIL: A PERSPECTIVE ON THE RISE AND ULTIMATE DEMISE OF GOOGLE

results displayed in the figure below, it should not be assumed that Google has intentionally taken a blow at its competitors. Literally, the democratic nature of the Internet (whose 'votes' were tallied by Google) has produced the ironic search results.

However, just because results are automatically generated does not mean that they are necessarily the most accurate, or that they are the most subjective and unbiased. The Internet as a whole is arguably considered to be slanted liberally⁹. Thus, if Google indexed solely on link popularity, it would follow logically that there would be a heavy liberal slant to political articles, for example. This must at least lead us to ask the question – how is Google biased and how or who does that bias benefit?

As it turns out, there is extensive research to support that Google is indeed biased – although its bias is paradoxically conservative¹⁰. How can one reconcile this seeming disparity in logic? A quick Google search for 'Google bias' returns, not surprisingly, a series of articles that indicate that Google is not in fact biased in any form. All of these articles suggest that Google is in fact a neutral entity, scouring the Internet and returning

⁹ http://64.233.161.104/search?q=cache:KVSAoCc_Cn8J:www.worldmagblog.com/MT/mt-comments.cgi%3Fentry_id%3D578+%22%2Bis+%2Bthe+internet+conservative+%2Bor+liberal%22&hl=en&client=firefox-a, "World Magazine Blog: Comment on A Media Revolution," January 2004
¹⁰ <http://ojr.org/ojr/technology/1095977436.php>, "OJR article: Balancing Act: How News Portals Serve Up Political Stories," November 2004

what is a collective body of knowledge, ranked by a source's popularity. A comparative search for the same terms via Yahoo returns ten completely different articles – most of which seem to agree that there is, or is a possibility of, conservative bias. I propose that such a disparity is no accident.

At the same time, Google is considered to be the bastion of Internet omnipotence. A search for politically valenced terms may indeed lead to politically motivated results regardless of the search method, but a search for some neutral phrase such as '1993 honda civic price', whose use would be more concrete and less objective would return the desired factual information.

WHAT'S THE DIFFERENCE?

One cause for concern with Google is that the spirit it embodies is that of such strong superiority and seeming infallibility that users will blindly trust information on any topic. It is arguable that a proportion of consumers of any medium will do the same thing, but there is one significant difference: most media outlets purport fairness without actually being fair at the cost of credibility. Presumably, few liberals would routinely watch Fox News because they would turn away from the clear bias injected by the nuances of a newscast. The "fair and balanced" tag is not taken seriously by those who disagree with conservative agendas. Google, on the hand, is generally perceived to be unbiased. When a

DON'T BE EVIL: A PERSPECTIVE ON THE RISE AND ULTIMATE DEMISE OF GOOGLE

human relates news articles there is an assumed understanding of some hidden agenda, or at the very least of the most rudimentary motivation. A computer cannot be motivated – but its programmers can.

Even if we assume a lack of said bias, Google is still worrisome solely because of its dominance. No matter how ‘neutral’ a source is considered to be, no one would dare suggest that all other news and information channels should be abandoned in favor of Google. Yet, in its peak in early 2004, Google search technology was responsible for generating approximately 80% of the referrers to external websites⁷. Consider here that this figure includes searches performed by other ‘brands’ of engines that simply license the Google technology or reuse their results. The 80% figure cited previously dropped sharply by spring of 2004 because of Yahoo’s decision to abandon the Google engine.

This is perhaps one of the most frightening aspects of the Google technology – it’s presence behind the scenes, in the shadows of server rooms in Silicon Valley. If Yahoo, AOL Search, CNN and Google are all one in the same (which, for a time, they were), then there is no true competition⁷. In the purest sense of the word ‘monopoly’ Google is not a monopoly, because it does not engage in anti-competitive acts. They argue they are simply the best and thus even their competitors license their technology. Google wishes it were only that simple.

Consider that Google, which can be argued to be slanted conservatively, is partnered with a well known and openly left slanted media giant such as CNN (owned by Time Warner, Inc., also the owners of AOL) – does the sum of the two simply cancel each other out? Does releasing one right wing article per left wing article keep the news Internet news sources balanced? More concretely, was this the intent of the Internet? The Internet was supposed to be an ‘open club,’ but how open can it be when two of the largest media outlets – both liberal and conservative – work together to bring the same information?

For that matter, is Google bringing us all the information it can? Take the recent controversy surrounding the torturing of prisoners at Abu Ghraib prison. One should expect that Google would automatically index photos that it sees frequently, without heavily manipulating image search results (it’s all automatic, remember). A Yahoo image search for ‘abu ghraib torture’ returned 317 of the same grisly photos shown *ad nauseum* on television around the time news of the scandal broke. Astoundingly, even with their ‘Safesearch’ obscene content filter disabled, a Google image search for the same terms return only two photos. This, certainly, is not a democratic representation of the content on the Internet, especially considering the context: the Abu Ghraib scandal was largely publicized, and the photos were shown in a variety of media. To suggest that Google simply did not find any

DON'T BE EVIL: A PERSPECTIVE ON THE RISE AND ULTIMATE DEMISE OF GOOGLE

websites that contained those same photos is preposterous.

Now recall that link popularity is used extensively to rank sites on their relevance. It is patently obvious that large news outlets such as CNN and Fox News will see more links, and thus will be considered more relevant. This, arguably, ruins the 'open' nature of the Internet. If all data mining attempts point back to the same sites, the greater proportion of the Internet gets ignored. Supporters of Google say that if the everyday freelance web developer created some news site, it simply wouldn't be relevant enough to warrant high ranks with Google. On the other hand, if Google heavily represents a miniscule number of actual different major news sources, then it becomes virtually impossible for any other piece of original information to disseminate to readers.

"The cure for cancer might already be on the web somewhere, but if it's on a new site, you won't find it."

-Daniel Brandt, founder & president, Public Information Research, Inc.

The Internet was designed to be a truly egalitarian system in which anyone could exchange ruminations in a free marketplace of ideas. The PageRank system that made Google popular completely strips the Internet of this freedom by reinforcing those coming into the Internet with the greatest name recognition and greatest bankrolls, making it easy for them to work their way to the highly coveted tops of rankings. Individual web developers will be hard

pressed to create much of a buzz, except maybe amongst close friends and family. These developers will simply not be able to generate the number of back links necessary to muscle to the top of the listings.

PageRank becomes more concerning with a simple comparison to affirmative action. While many people disagree with the principle of affirmative action, those that support it do so because it theoretically gives the less privileged more opportunities. Those who disagree with it consider it to be reverse discrimination. No one in their right mind would suggest that any kind of reverse affirmative action should take place – providing more opportunities for the wealthy and powerful. In a way, Google's PageRank technology does just this. Fresh new web sites are left with virtually no search engine traffic while well established commercial sites dominate. A new web developer could publish their work for over a year, constantly trying to coax other publishers to link to them, only to find that they receive a miniscule number of hits. Sites with even moderate PageRanks may receive well over 10,000 hits per day¹¹.

Is Google trying to stifle amateur web publishers? With all the advances Google has given to the field of search technology, are they still unable to implement any kind of content analysis? Can a computer not be trained to

¹¹ <http://www.google-watch.org/pagerank.html>, "PageRank: Google's Original Sin," November 2004

DON'T BE EVIL: A PERSPECTIVE ON THE RISE AND ULTIMATE DEMISE OF GOOGLE

search for well cited, grammatically correct, content rich sites? Although difficult, it seems that such a technology would be worth pursuing. Without some kind of bottom-up topology, Google will continue to propagate information from the same websites.

DOES IT EVER END?

As with any real system, Google too must be finite. In 2003, over a year before Google's IPO, there was widespread speculation amongst web developers that something had gone wrong at Google. It seemed that the usual monthly visit from the Google web spiders did not come. More interestingly, it was apparent that Google had reverted to a backup of its index that was roughly two months old. Observant programmers noted that Google's total index size (the number of unique pages that Google made searchable) was approaching, and beginning to exceed 4 billion. As it turns out, when Google was being coded in the mid 1990s, an Internet with four billion pages seemed to be quite a large number. Additionally, from a performance perspective, programmers are taught that optimization is critical to creating good code. For these reasons, it seems obvious why Google broke – Google was designed during a different era, one in which a four byte “document ID” would not only suffice, but would provide generous room for expansion. Every item on Google is indexed by such an ID.

As Google approached four billion pages, developers must have been getting worried that their code would eventually cease to work. This is because the ANSI C programming language uses certain ‘variable types’ that are salient to varying sizes of numbers. For memory efficiency purposes, variables can be Boolean, or at the other end of the scale, incredibly large. The largest variable type natively supported in the ANSI C language is the ‘unsigned long integer’, with a maximum value of precisely 4,294,967,296. So why not simply make the switch to some superior numbering system? To do so would require the change of every document ID from a four byte to five byte identifier. Thus, the space required to index every word in every document would increase by roughly 20%¹². Google simply was unable to complete an upgrade task of that scale without losing some functionality. Today, Google indexes approximately 8 billion pages – indicating that they have, at the very least, added a fifth bit to their document ID field.

Considering that Google is by far the largest supercomputer in the world according to estimates⁷, making a sudden 20% increase in size would not be a simple task. Alas, they were successful in migrating to the more expansive database system. So can anything stop Google?

¹² <http://www.google-watch.org/broken.html>, “Is Google Broken?,” November 2004

DON'T BE EVIL: A PERSPECTIVE ON THE RISE AND ULTIMATE DEMISE OF GOOGLE

The way it seems, Google's competitors are focusing more on emulating Google than on innovating. Many of the elements that made Google popular – such as its minimalist interface without unnecessary flashy graphics and distracting visual advertisements. This could be seen as an overall improvement in the quality of service afforded on the Internet. Search engine startups such as Exalead purport themselves to be 'unique' in some way but fail to distinguish themselves in any single category.

CONCLUSIONS

Given Google's current grasp on the majority of searches worldwide, it seems that unless another company develops a truly miraculous search technology, Google's heavy market share and generous financial backing make it a firm competitor. The solution to Google, it seems, is to not use it. It is for this reason that I view Google's current position to be detrimental to the Internet as a medium. The entire point of the distributed nonhierarchical structure was to prevent a single entity such as Google from dominating the traffic on the network.

At the same time, the Internet is now almost a basic utility to some people. If broadband-to-the-curb initiatives are undertaken in Congress, and the Internet does become a public utility, will Google be subject to regulation as any other utility provider would?

Is it possible that Google's popularity is just a sign of the times? America Online was one seen as the conglomerate Internet dominator, but is now the ridicule of many who use computers or the Internet professionally. Perhaps the most telling sign that Google is a little bit too intimately involved in our daily life. Below, a slide used while explaining this assignment to our class.

Using Sources

- ◆ **Don't hand in other people's work:**
 - ◆ You won't learn from the assignment.
 - ◆ You will feel guilty.
 - ◆ Google will help you get caught.
- ◆ **It is a good idea to use sources**
 - ◆ Do a Google search to get the lay of the land
 - ◆ Try Amazon for books on the topic
 - ◆ Look up the original sources
- ◆ **It is fine to cite sources; it is not fine to pretend you had none**

7

When Google is omnipresent in this state, I believe it becomes a matter requiring involuntary regulation – even if only to ensure that Google does not abuse its dominance by engaging in unethical behaviors such as censorship and bias.

Peter Ubriaco is a third year student at Rensselaer Polytechnic Institute, pursuing a dual degree in Chemical Engineering and Psychology. He has recently worked professionally developing organizational computer security policies and developing custom web content management systems. His personal homepage has a PageRank of 4. He can be reached by email at ubriap@rpi.edu.